DOCUMENT RESUME

ED 178 566 TM 009 710

AUTHOR	Merz, William R.; Grossen, Neal E.
TITLE	An Empirical Investigaiton of Six Methods for Examing Test Item Bias. Final Report.
INSTITUTION	Sacramento State Univ. Foundation, Calif.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	[78]
GRANT	NIE-6-78-0067
NOTE	59p.
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Complexity Level; *Evaluation Methods; Factor Analysis; *Item Analysis; Mathematical Models; Simulation; Statistical Analysis; *Test Bias; *Test Items; Test Reliability; Test Results; Test Validity
IDENTIFIERS	Chi Square; Item Characteristic Curve Theory; One Parameter Model; Point Biserial Correlation; Three Parameter Model

ABSTRACT

Six approaches to assessing test item bias were examined: transformed item difficulty, point biserial correlations, chi-square, factor analysis, one parameter item characteristic curve, and three parameter item characteristic curve. Data sets for analysis were generated by a Monte Carlo technique based on the three parameter model; thus, four parameters were controlled: total score distributions, item difficulties, item discriminations, and guessing. Only the difficulty parameter was biased. Results indicated that transformed item difficulty had highest correlations with generated bias. The three-parameter item characteristic curve and the one parameter item characteristic curve were next highest, followed by factor analysis, then chi-square. Point biserial correlation functioned erratically. Results of the analysis were compared and recommendations were made on the use of each method. (Author/MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED178566

FOUNDATION
of
CALIFORNIA STATE UNIVERSITY, SACRAMENTO

6000 J STREET • SACRAMENTO, CALIFORNIA 95819 • (916) 454-6635

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

William R. Merz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FINAL REPORT

Grant NIE-6-78-0067

An Empirical Investigation of Six Methods for Examining Test Item Bias

Co-Investigators

William R. Merz, Ph.D.
Department of Behavioral Sciences in Education

Neal E. Grossen, Ph.D.
Department of Psychology

Research Assistants

Mary Lynn Groome, MA
William R. Groome, MA

TM009 710

This research was sponsored with funds from the National Institute of Education, Grant NIE-6-78-0067. The views and opinions expressed in it are those of the authors' and do not necessarily reflect those of the National Institute of Education or the Department of Health, Education, and Welfare.

Special thanks to Lawrence M. Rudner, Ph.D. who served as consultant to the project for his assistance with programming, data analysis, and concept development and to Reginald A.H. Goodfellow, Ph.D. for his editing and critical review of the final report.

ABSTRACT

The purpose of this investigation was to examine six approaches for assessing test item bias. The methods employed were Transformed Item Difficulty, Point Biserial Correlations, Chi-Square, Factor Analysis, One Parameter Item Characteristic Curve, and Three Parameter Item Characteristic Curve. Data sets for analysis were generated by a Monte Carlo technique based on the three parameter model; thus four parameters were controlled: total score distributions, item difficulties, item discriminations, and guessing. Only the difficulty parameter was biased. Results indicated that Transformed Item Difficulty had highest correlations with generated bias. The Three Parameter Item Characteristic Curve and the One Parameter Item Characteristic Curve were next highest. Factor Analysis then Chi-square followed next. Point Biserial correlation functioned erratically. Results of the analysis are compared and recommendations on the use of each method are presented.

TABLE OF CONTENTS

	PAGE
Abstract	1
List of Tables	3
List of Figures	4
SECTION	
BACKGROUND	5
Purpose	5
Recent Considerations of Bias in Testing	5
What Is Bias?	7
The General Situation	7
Statistical Approaches	7
Validity and Reliability	10
Ability and Bias	12
METHODS OF ASSESSING TEST ITEM BIAS	14
Transformed Item Difficulty	14
Point Biserial Correlation	17
Chi-Square	18
Factor Analysis	20
Item Characteristic Curve	22
METHOD	32
Overview	32
Data Generation	32
Methods of Analysis	35
Agreement of Methods	36
RESULTS	38
Characteristics of Data Sets	38
Comparison of Methods	41
ICC	41
Point Biserial Correlation	43
Transformed Item Difficulties	43
Chi-Square	43
Factor Analysis	44
Accuracy of Identification	44
CONCLUSIONS AND RECOMMENDATIONS	46
Summary	46
Conclusions	46
REFERENCES	50

LIST OF TABLES

NUMBER		PAGE
1	Summary of Data Sets Generated for Analysis . .	38
2	Test Statistics for Two Groups of 1000 Examinees .	40
3	Correlations between the Degree of Bias Identified by each Method and the Amount Generated with BIGEN	42
4	Comparisons of Number of Items Identified as Biased and Number of Biased Items Correctly Identified	45
5	Comparison of Methods	48

LIST OF FIGURES

		PAGE
Figure 1:	A hypothetical transformed p-values regression scatterplot.	16
Figure 2:	Two hypothetical response distributions . .	25
Figure 3:	A hypothetical ICC	27
Figure 4:	Bias Solely in a_g	28
Figure 5:	Bias Solely in b_g	28
Figure 6:	Bias in Both a_g and b_g	28

BACKGROUND

Purpose

The purpose of this investigation was to examine six approaches for assessing test item bias by analyzing common data sets with four prespecified parameters: total score distributions, item difficulties, item discriminations, and guessing. Results of the analyses are compared in light of the definitions; assumptions; and limitations, both logical and mathematical, which each model imposes. Recommendations on the use of each method are presented.

Recent Considerations of Bias in Testing

The question of bias in testing is as old as the concept of testing itself. The more dependent our educational systems become on using test scores for evaluation and diagnosis, the more important it is for test developers and administrators to assess and analyze test scores for bias. The National Institute of Education (NIE), recognizing the importance and timeliness of this topic, sponsored a three-day conference in December 1975, at Anapolis, Maryland. During that conference, experts in the field of testing were able to discuss approaches for constructing less biased tests and to

compare the results of "debiasing tests."

The United States Office of Education (USOE) recognized the difficulty experienced by minority children in programs funded for the improvement of academic attainment. An invitational conference was sponsored in May 1976, to examine achievement testing of disadvantaged and minority students. Some of the statistical techniques examined in this study were presented at or were discussed during that conference; a number of considerations other than statistical methods were discussed, as well.

Since the two conferences, substantial progress in the application of item characteristic curve (ICC) approaches has been made. ICC approaches relate an individual's ability to the probability of getting an item correct. However, rather than bringing clarity to the issue of test bias, the introduction of ICC methods has added to the confusion.

An entire issue of the Journal of Educational Measurement (JEM) was devoted to the fair use of test scores (1976). Latent trait approaches to test construction, including application to examination of test items for bias, were presented in recent issues of JEM (1977) and the Review of Educational Research (RER, 1978).

Professional associations interested in testing have addressed the issue of bias in testing. Sessions at the annual meetings of the American Educational Research Association (AERA), the National Council on Measurement in Educational (NCME), and the American Psychological Association (APA) in 1978 and 1979 were devoted to examining test items for bias. Educational Testing Service (ETS,

1979) has included an entire segment on bias in testing in materials developed for the California State Department of Education (CSDE), on construction of tests for mandatory competency testing programs. Thus, it is evident that interest in the topic of bias in testing has been growing.

What Is Bias?

The General Situation. The assertion that a test is biased is based on an individual's dissatisfaction with her/his test score or with decisions made about her/him based on test scores. First, an individual has taken a set of items which he/she believes to have measured her/his attainment inaccurately. Second, the individual is a member of a specific group which, on the average, attains scores different from another identifiable group. Often, the assertion made is that a test is invalid for a specific individual because he sheis a member of a minority group. The situation is that as a member of the minority group, the individual obtains a score lower than he/she feels represents her/his true ability or achievement.

Statistical Approaches. In such a situation the definition of bias is at issue. Is bias created during the actual construction of a test or is bias involved in the unfair use of test results? According to one definition, *item bias* is judged without outside criteria; it is measured by examining test items in relation to total test scores, item difficulties, or other criteria specific to the

individual test itself (Merz, 1978). In contrast with that definition, *test bias* (also known as *fairness*) is determined by whether or not similar predicitons and/or descriptions are obtained for each group tested (Potthoff, 1966; Merz, 1978; Rudner, 1977d; Merz and Rudner, 1978).

The present investigators contend that *test bias* or *fairness* is an issue which can be judged only after a test instrument has been constructed and is in use. At that point whether or not selection favors members of one group over members of another group can be observed. *Item bias*, however, must be examined during the construction of the test.

The clearest discussion of the problem of defining and assessing *fairness* is found in Petersen and Novick's critique (1976), which examines seven models used to analyze fairness in testing. All of these models concentrate on fair test use rather than on unbiased measurement.

The issue of test *item bias* is equally problematic, and attempts to investigate it systematically are relatively recent. The six best known procedures (analysis of variance, transformed item difficulties, item-total score correlations, multivariate analysis, chi-square, and response foil procedures) have been described by Merz (1978). Although all but one of these six procedures begin with zero-one scores, that is, right-wrong item scores, each approaches the analysis with different methodologies. Rudner's important

review (1976f) of techniques for assessing item bias considers 60 to 70 references and discusses six procedures: analysis of variance, transformed item difficulty, factor analysis, item characteristic curve, chi-square, and foil response analysis.

Current statistical methods present many conflicting procedures (Merz and Rudner, 1978). As Merz (1978) stated, "finding adequate criterion measures confounds the logical procedure with the statistical methodology". These problems are compounded because criterion data, "are difficult and expensive to obtain even when agreement on the adequacy of the criteria can be reached" (Merz, 1978).

Since the preparation of Merz's review (1978) additional work on test item bias has been completed. The entire summer 1977 issue of the Journal of Educational Measurement was devoted to an explanation of latent trait (ICC) models. Hambleton, et al. (1978) have presented an excellent review of latent trait (ICC) approaches. Weiss (1977) edited proceedings of a seminar on computerized adaptive testing; the method of selecting items for presentation is ICC theory. Rudner has presented at least three additional clear explanations of methods used to examine a test item pool for bias (1977a, 1977d, and 1979a). In these presentations he has discussed both the one and three parameter logistic models.

These additional examinations of test bias have tended to confuse methodological issues for the practioner who has not seen a comparison of approaches using a single data set. In fact, only

brief beginnings at comparing methods and describing conditions for using different approaches have been made. Bourne (1977) analyzed a contrived data set using four traditional models. Rudner (1977b) examined two data sets using a transformed item difficulty approach, an adjusted chi-square approach, an ICC approach, and a factor score approach. William Groome (1977) and Mary Lynn Groome (1977) compared the transformed item difficulty approach to the three parameter model with which hypothetical data were generated. This was done under uniform and normal distribution of item parameters. Ironson (1978) compared several methods. However, these studies presented only partial comparisons because the essential parameters underlying a distribution of scores were incompletely varied. In order to investigate adequately the limits of analysis techniques, the present investigators contend that the parameters of underlying ability, item difficulty, item discrimination, and guessing must be systematically varied in order to compare methods. Rudner, Getson, and Knight (1979b) and the present investigators have begun such studies.

Validity and Reliability. When a person claims that test results are biased against her/him, two questions must be addressed. First, is the sample of tasks presented a valid representation of the domain of tasks to which the total score is to be generalized? That is, how well do the tasks represent the domain to be tested? Second, is the individual's score a reliable representation of

her/his performance on the tasks in the test which have been sampled representatively from the domain of interest. Put another way, is the test score a reliable representation of this individual's performance?

The issues are not simple. However, in order to clarify what is being discussed it is necessary to distinguish between the two sampling domains. First, to address the issue of validity one must examine how well the set of tasks on a test sample the domain to be tested. In other words, the issue of validity approaches the sampling of items included on a test. Reliability, on the other hand, addresses how representative of typical performance an individual's performance is; that is, how typical this performance is of a person's usual performance in the domain being tested.

In reviewing classical test theory one sees that the equation below

$$X = T + \varepsilon$$

where: X is an individual's obtained score,
T is an individual's true score, and
ε is error associated with measurement.

addresses these issues in this way. When one asks how well a test or set of tasks samples the underlying construct to be measured, one

asks how well the test items or the tasks sample the universe represented by the construct. When one asks how representative an individual's test score is of her/his typical performance, one asks how free from error of measurement observation of performance is or how typical of the domain of the individual's performance her/his score is.

The issues are not as clear as this categorization seems to indicate. When one asks whether or not an examinee's performance is representative of the construct to which a generalization is being made, one confounds the issues validity and reliability. It is exactly this type of confounding which has led to statements that specific tests or test situations are invalid for members of specific groups. The classification system describing sources unreliability developed by Thorndike (1951) and reiterated by Stanley (1971) may clarify the situation. Usually variation of performance within a group on any set of tasks is larger than variation between groups. However, the ratio of variance between groups to total variance may be larger than one would desire. It is exactly this phenomenon which must be examined to determine whether or not a specific task or a test item is biased.

Ability and Bias. It appears that two conditions of test item bias exist. The first condition exists when two groups have equal ability and specific tasks or items are attained correctly by a greater proportion of one group than an other. The second,

when the groups have different levels of ability and items or tasks are attained correctly by different proportions in each group. The second situation poses quite an interesting problem. How it is possible to discern whether the underlying ability of the group or some specific feature about the way a task or question was presented has created the difference in proportion of correct response? In this situation, group ability and differential proportion of correct responses due to characteristics specific to the item become confounded.

A central issue, then, in dealing with test item bias is accounting for differences in ability. In order to address this issue, a definition of test bias coming from ICC theory best separates the issues. This definition is *bias is a differential probability of correct response given equal ability*. Under this definition a task in a performance exercise or item in a test is biased, if individuals of the same ability level have different probabilities of correct response because of their group membership. This definition is appealing because it seems to avoid the thorny theoretical issues discussed earlier. One no longer needs to worry about the underlying construct being measured when one has a reasonable estimate of ability on that construct. This, though, is the very weakness of the definition. Rarely is it possible in examining test item bias to have a measure of ability independent of test performance being examined.

METHODS OF ASSESSING TEST ITEM BIAS

Six methods of assessing test item bias are examined in this paper. They are:

Transformed Item Difficulty,
Point Biserial Correlation,
Chi-Square,
Factor Analysis,
One Parameter Item Characteristic Curve, and
Three Parameter Item Characteristic Curve.

Each is considered below; definitions, assumptions, limitations, and adjustment for different levels of ability are treated.

Transformed Item Difficulty

This approach probably was described first by Thurstone (1925) as part of his method of absolute scaling. It may be the best known of the six methods because it has been used by many investigators: Angoff (1972); Green and Draper (1972); Angoff and Modu (1973);

Jensen (1973), Angoff and Ford (1974); Echternacht (1974); Strassberg-Rosenberg and Donlon (1975); Hicks and others (1976); and Rudner (1978).

The procedure is to transform item p-values (percent of correct response) to z values for each group independently; each group's average p-value and standard deviation of p-values for the entire test is used.

$$z = \frac{p_i - p.}{S_p}$$

where: p_i is the percent of pupils passing item i,

$p.$ is the mean percent correct for the whole test, and

S_p is the standard deviation of the percent correct for the whole test.

These z values are converted to delta values (Δ) independently for each group using the formula $\Delta = 13 + 4Z$, where the mean of Δ's is 13 and the standard deviation is 4. Each pair of Δ's is plotted on a bivariate graph. A line 45° to the abscissa is plotted through the origin to indicate absence of bias. Confidence bands are drawn around this line. Item pairs which have plots falling outside the bands are considered biased. This is depicted in Figure 1.

A variety of approaches to setting bands around the line have been suggested. Echternacht (1974) employed hypothetical lines based on the sample size and significance level. Another

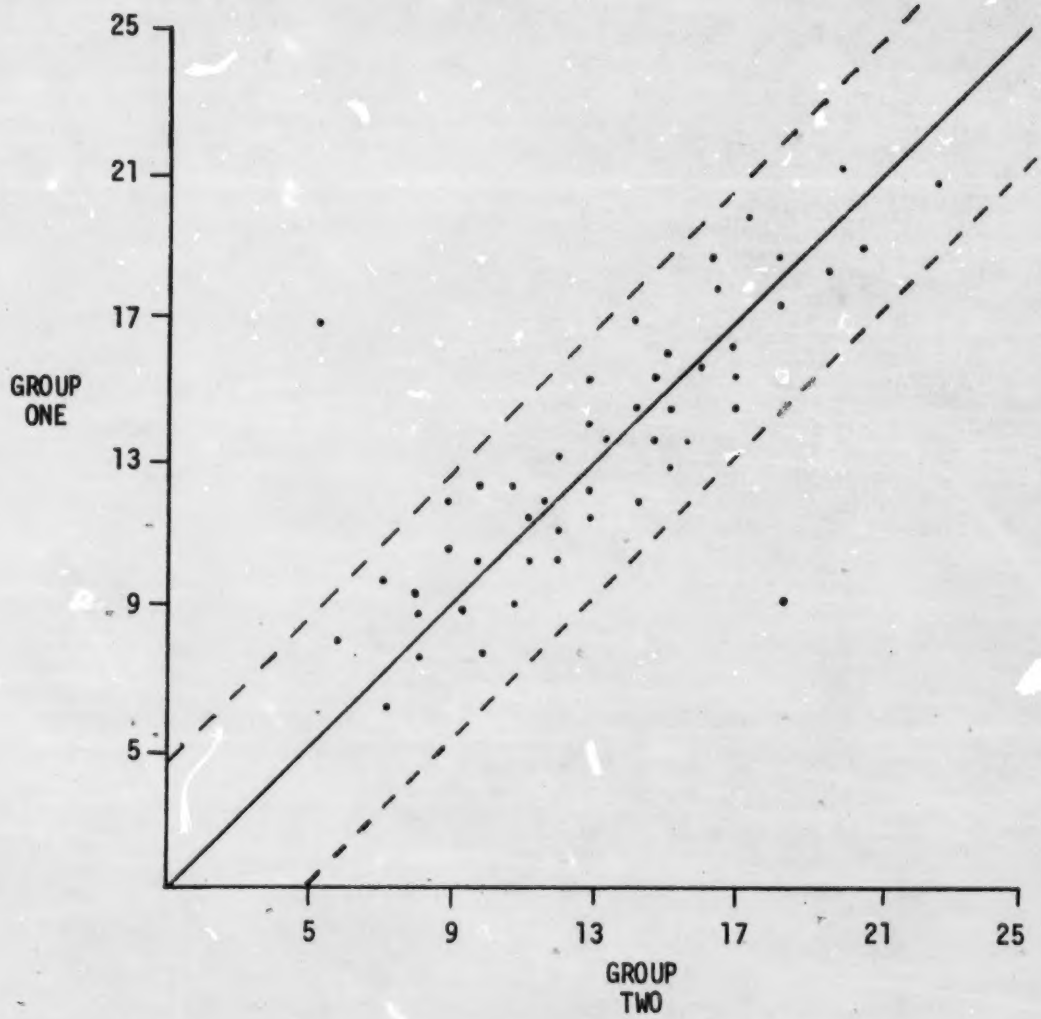

Figure 1. A hypothetical transformed p-values regression scatterplot.

approach involved computing the standard deviation of residuals, distances from the regression line, and labeling as biased those items which deviate more than 1.5 standard deviation units (Strassberg-Rosenberg, 1975 and Hicks and others, 1976). Rudner (1978) suggested a fixed distance of .75 z score units. Still another approach has been to use a multiple of the standard error of measurement.

The TID approach defines item bias as item by group interaction; that is, the magnitudes of item difficulty deviations from the average test difficulty differ from group to group by an amount greater than that specified by the investigator. This method adjusts for differences in group total score means, which may be considered a reflection of a group's ability, by computing z scores independently for each group. The technique assumes that relative differences in difficulty reflect bias rather than differences in ability. No outside criterion is used. And the method is applied to only two groups at a time.

Point Biserial Correlation

This approach examines the correlation between item performance and total score. It has been employed by Green (1971, 1973, 1976), Green and Draper (1972), and Ozenne, Van Gelder, and Cohen (1974). Differences between pairs of correlations for each item are computed for two groups at a time. Differences which exceed

a prespecified value are considered biased. Rudner, Getson, and Knight (1979a) considered this a modification of the transformed item difficulty approach citing Green and Draper (1972) who saw this method as estimating regression in such a way that item difficulties could be formed "in a manner analogous to the way in which adjusted means are formed in the analysis of covariance (1972, p. 16)." The adjustment to differing ability levels or mean values is made automatically if one considers the correlation approaches as Green and Draper do. However, both item discimination and item difficulty influence item location on the regression plot in this procedure; whereas, with the transformed item difficulty p-values alone determine the location. Because the individual items are scored right or wrong, the magnitude of the correlation coefficient is minimized. With this particular approach total score enters directly into the computation and may adversely affect the assessment of bias because the total score is an aggregation of right responses including the items which may be biased.

Chi-Square

Scheuneman (1975, 1976) developed a technique which sought to determine whether examinees of the same ability level had the same probability of a correct response regardless of group membership. This was done by blocking each group into three to five classes based on observed total scores and comparing the

proportions of individuals within each class responding correctly. An item is considered unbiased if, for all individuals in the same total score interval, the proportion of correct responses is the same for both groups under consideration. A modified chi-square is used to estimate the probabilty that the item is unbiased. The expected values for each cell (E_{ij}) are obtained by multiplying (1) the proportion of all examinees within total scores within interval j responding correctly to item i by (2) the number of examinees within the cell. Algebraically, that is,

$$E_{ij} = \frac{O._j}{N._j} N_{ij}$$

where: $O._j$ is the number of examinees in total score interval j responding correctly,

$N._j$ is the total number of examinees in total score interval j, and

N_{ij} is the total number of examinees in group i and score interval j.

This method may be considered an item characteristic curve approach which utilizes item difficulty. Bias is defined as different proportions passing an item at given levels. Here, the ability level is reflected in total score, and proportion correct should be identical for each level of ability. By blocking groups

separately into ability levels, this method adjusts for differences in mean level of attainment. No outside criterion is used, and the procedure may be applied to more than two groups at a time.

Factor Analysis

In factor analysis, underlying dimensions or traits are hypothesized, and the correlations of each variable with the hypothesized factors are computed. In an achievement test, each item is treated as a variable. Such an analysis could be conducted twice using examinees from two different cultural backgrounds. Ideally, the two separate groups of examinees would yield similar sets of item-trait correlations or factor loadings. Different sets of factor loadings would indicate that the two groups did not respond to the items in the same way. Such a test item would be considered biased because it would appear to measure different traits across groups. The items exhibiting the most bias would then be those with the largest differences in factor loading.

The general model for this type of factor analysis is

$$F = Y_z R^{-1} A + E$$

where: Yz is a vector of subject responses in z score form
A is a matrix of factor loadings
F is a subject vector of factor scores
E is a vector of residual or error terms
R is the variable by variable correlation matrix

Green and Draper (1972) and Green (1976) suggested an inter-group factor analysis model based on the inter-battery factor analysis approach offered by Tucker (1958). In this inter-group model, the item variance is partitioned into (1) factors common to each subgroup, (2) factors specific to subgroups, and (3) residual or error variance. With this model one can determine the proportion of item variance accounted for by a given subgroup. An item, then, is unbiased when the proportion of variance accounted for by a specific group is small and biased if a large proportion of variance is attributable to a factor specific to a subgroup.

Merz (1973, 1976) developed an alternate approach which incorporates factor scores and analysis of variance. The item intercorrelation matrix is computed for subjects pooled across groups. The matrix is reduced with Principal Components Analysis, employing the Scree technique to determine the number of factors to be extracted. The factor matrix is then rotated orthogonally to approach simple structure, and factor scores are derived from the rotated matrix. An analysis of variance is then conducted on each set of factor scores using multiple group memberships as independent variables and factor scores for each vector as the dependent variable. Items are identified as biased if they have a major loading on a factor with a significant F ratio on a main effect or on an interaction. Thus, differences in average factor score for a specific group indicate bias. Because item intercorrelations are analyzed with

this technique, mean total score differences should not affect the analysis.

Another approach suggested by Merz (1976) is to enter group membership into the item intercorrelation matrix as dummy variables. These dummy variables are intercorrelated along with item performance. The intercorrelation matrix is reduced with parincipal components analysis as in the method above. After orthogonal rotation, factor loadings are inspected; items which have major loadings on the same variable as group membership variables are considered biased. Again mean differences in total score do not enter into computation. The magnitude of the correlations in the matrix are affected by the correct-incorrect scoring. All three factor analysis methods are latent trait approaches as are the item characteristic curve approaches.

Item Characteristic Curve

Recently latent trait theory has been used to identify biased items (Green and Draper, 1972; Lord, 1977; Rudner, 1977a; Pine, 1977; Scheuneman, 1975 and 1976; Durovic, 1975; and Wright, Mead, and Draba, 1978). Latent traits refer to examinee characteristics, estimating scores for examinees on these traits, and using these scores to explain test performance (Lord and Novick, 1974). Latent trait theory deals with observable scores and their relationship to the unobservable traits which are represented by performance on

the test. Hambleton and others (1978) stated that there are three important concepts which must be considered in latent trait theory: (1) dimensionality, (2) local independence, and (3) item characteristic curves.

Dimensionality refers to the assumption that a test is unidimensional or measures a single underlying ability. Local independence means that a person's answering one item correctly is independent of or not affected by answering another item correctly. Item characteristic curves refer to the hypothetical distribution relating ability (θ_i) to the probability of obtaining a correct response to an item $P(u_g = 1|\theta_i)$ in terms of the characteristics of the item.

While various models (Lord, 1952; Rasch, 1960; Birnbaum, 1968; and Urry, 1970) differ in the number of item parameters or characteristics considered; they all define the item parameter(s) independently of the sample examined. This property has led to the development of some interesting applications in test development, tailored testing, and equating.

In an early study, Green and Draper (1972) had used observed total scores as estimates of examinees' abilities (θ_i's) and the proportions of examinees responding correctly at each total score level as estimates of the probability of a correct response given $\theta_i[P(u_g = 1|\theta_i)]$. Their procedure called for plotting separately for each group estimated ICC's and comparing the plots.

By this and other latent trait theory approaches, an item is unbiased if examinees of the same ability level, but of different group memberships, have equal probabilities of responding correctly. That is, an item is unbiased if the estimated ICC's obtained from the various culture groups are identical. As an example of a biased item, consider the two hypothetical curves shown in Figure 2. These curves are based on responses by two different groups to the same item. Total observed scores are used as estimates at θ_i and proportions of examinees responding correctly are used as estimates of $P(u_g = 1 \mid \theta_i)$. The curves are not identical, since the location parameters for the two curves are not equal. Such an item can be considered biased because examinees of the same ability level but from different groups do not have similar proportions of correct responses.

While this approach is appealing, total observed scores are direclty incorporated into estimating bias. Quantification of the degree of item bias is difficult; because visual inspection is used to identify a "very biased item."

A general, cumulative logistic model was proposed by Birnbaum (1968); it uses three item parameters:

a_g - an item discrimination index,

b_g - an item difficulty index, and

c_g - a pseudo-guessing parameter.

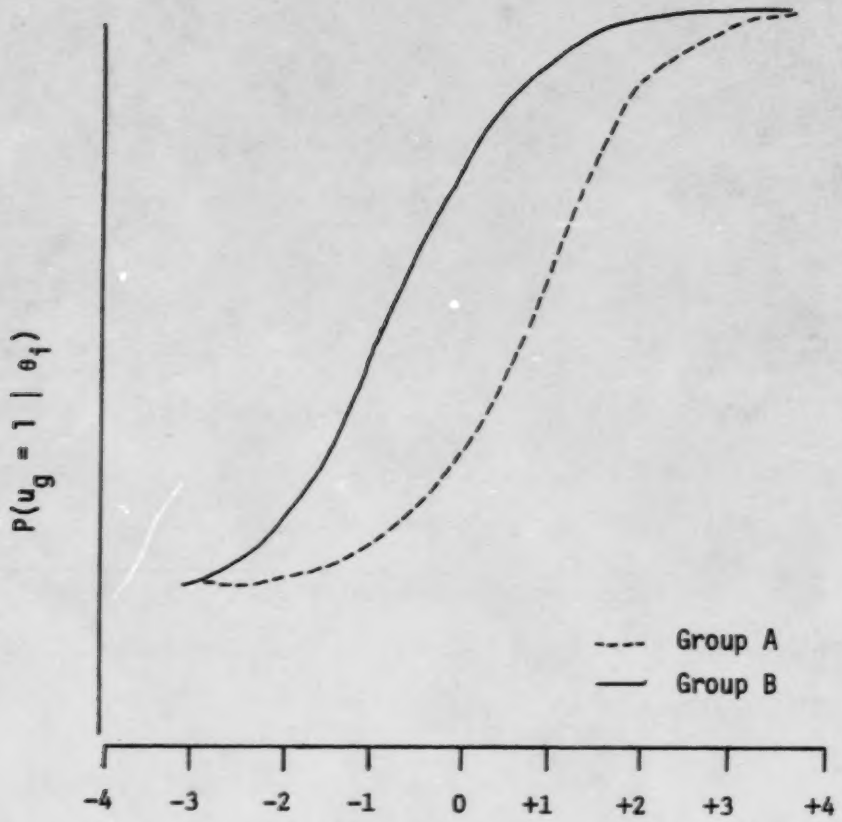

Figure 2. Two hypothetical response distributions

Using the notation $P(u_g = 1 \mid \theta_i)$ to represent the probability of a correct response to item g given an examinee of ability θ, Birnbaum's three parameter model may be stated as

$$P(u_g = 1 \mid \theta_i) = c_g + (1 - c_g)1 + \exp\{[-1.7a_g(\theta_i - b_g)]\}^{-1}$$

where: a_g is item discrimination,
b_g is item difficulty, and
c_g is an estimate of guessing

The relationship between $\theta_i (-\infty < \theta_i < \infty)$ and $P(u_g = 1 \mid \theta_i)$ is shown in Figure 3. Here $c_g = 0.2$, the lower asymptote; $b_g = 0.0$, the inflection point of the curve; and $a_g = 1.5$, a value which is directly related to the slope of the curve at b_g.

In Figure 4 the curves cross because the a_g parameters are not equal. In Figure 5 there is an area between curves because the b_g parameters are not equal.

In Figure 6 the curves are not parallel because the a_g parameters are not equal, and the inflection points are not at the same location because the b_g parameters are unequal. If these curves represent the same item, use the same metric for the ability scale, and are determined by separate group membership, then the item can be considered biased. That is, examinees of the same ability level but from different groups have unequal probabilities of responding correctly.

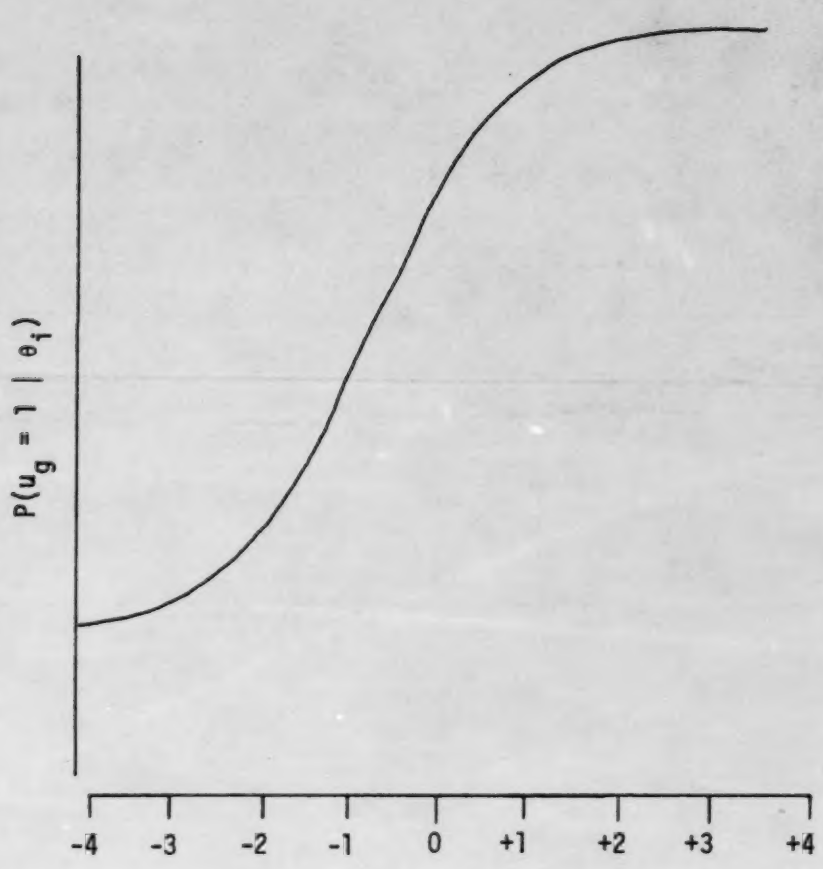

Figure 3. A hypothetical ICC

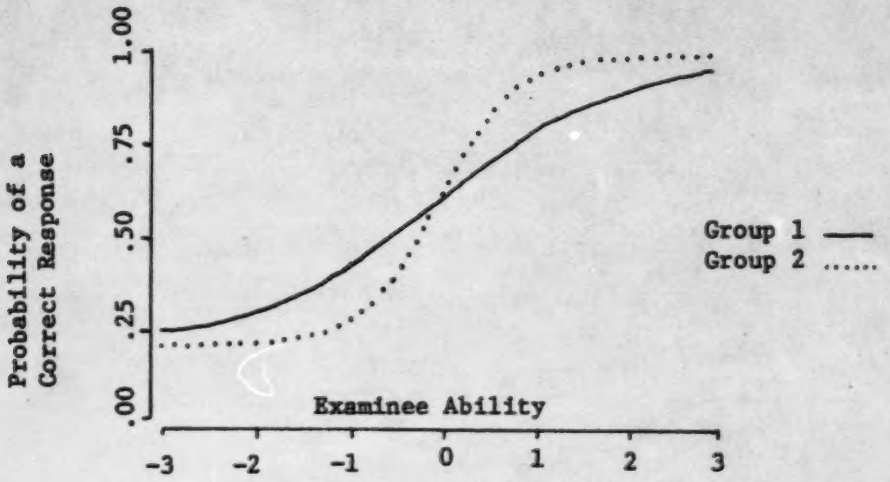

Figure 4. Bias Solely in a_g

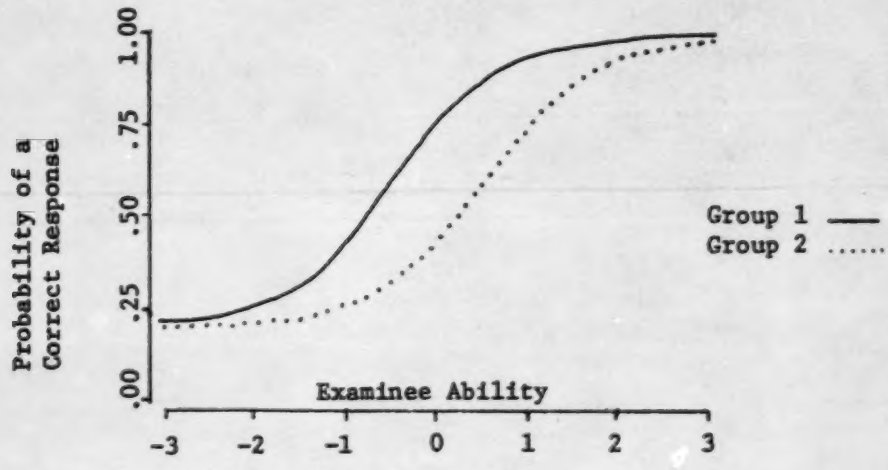

Figure 5. Bias Solely in b_g

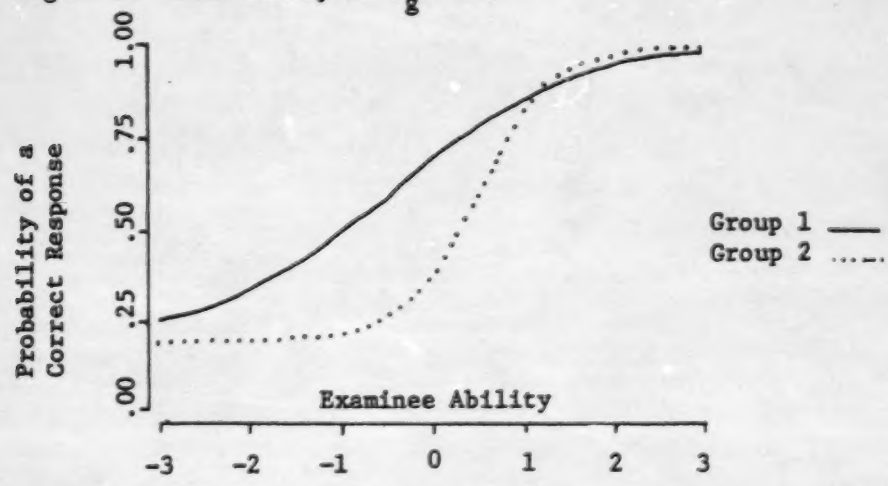

Figure 6. Bias in Both a_g and b_g

Rather than using total observed scores as estimates of θ_i and proportions as estimates for $P(u_g = 1 \mid \theta_i)$, more accurate values can be obtained using one of the recent methods of parameterization (Urry, 1975 or Wingersky and Lord, 1973). During parameterization, the metric used for the θ scale is defined by the ability variance in the sample of persons taking the test. Parameters obtained from two different examinee groups are compared by equating the obtained values. Lord and Novick (1974, Chapter 16.11) and Rudner (1977b) have accomplished this by computing the regressions of the parameter values based on one group of examinees against the parameter values based on another group of examinees. Equated ICC's will be identical when the measure

(1) is unidimensional,

(2) contains locally independent items, and

(3) has error free parameter estimates.

This assumes that accurate parameter estimates can usually be obtained from test performance, and answering one item correctly usually does not depend on answering another one correctly. Therefore, nonidentical ICC's would usually indicate that the test was not unidemensional; in other words, the items probably measures different traits across groups, what Rudner (1979b) calls bias, or traits other than those measured by the other items, what Rudner (1979b) calls inappropriateness.

The extent of bias could be evaluated by residuals from regression, Y' - Y. Rather than using residuals, distances such as those suggested by Angoff (1972) can be used.

Another approach is to adapt the method suggested by Green and Draper (1972), using the equated parameter values to plot ICC's for each group. This avoids making the scale depend on potentially biased total scores more accurately representing $P(u_g = 1 \mid \theta_i)$. However, visual inspection would still be necessary.

A more attractive approach was used by Lord (in press) for the three parameter model. He tested for a significant difference between equated ICC's. An asympotic significance test based on the summed variance - covariance matrices of the a_g and b_g parameter estimates was employed.

Rudner (1979b) preferred to compute the area between the two equated icc's. The size of the area would be low for relatively unbiased items and high for relatively biased ones. This area is defined by the equation

$$\phi_g = \int_{-\infty}^{\infty} [\,|\, P(u_g=1|\theta_i) - P'(u_g=1|\theta_i) \,|\,] \, d\theta$$

where $P(u_g=1|\theta_i)$ and $P'(u_g=1|\theta_i)$ define the equated icc's for the two groups.

It can be approximated with a computer by the equation

$$\phi_g = \sum_{-5.000}^{5.000} [\,|\, P(u_g=1|\theta_i) - P'(u_g=1|\theta_i) \,|\,] \,\Delta\theta$$

where $\Delta\theta = .005$.

Here bias is placed on a ratio scale. This approach overcomes the problems of visual inspection, and it simultaneously analyzes differences in item discrimination and difficulty.

A one parameter or characteristic model was proposed by Rasch (1960) and adapted by Wright (1977). The entire solution is based on this equation

$$\Pr\{X_{vi}/\beta_v,\delta_i\} = e^{X_{vi}(\beta_v-\delta_i)} / \{1+e^{(\beta_v-\delta_i)}\}.$$

This solution is a particular condition of Birnbaum's model where guessing and discrimination are held constant. Only the examinee's ability and the item difficulty enter. These are estimated by approximation. Equivalent results should be obtained by the Birnbaum and Rasch models when guessing and descrimination are held constant.

METHOD

Overview

This study compared six methods of assessing test item by analyzing data sets generated with prespecified total score means and standard deviations, item discriminations, item difficulties, and guessing estimates. Data with known properties was generated for analysis rather then collecting actual data to control parameters known to influence analysis. The point of this study was to compare the function of the statistical techniques rather than to determine the degree of bias in an actual set of data. In this way the limits of the methods could be tested.

Data Generation

Item data was generated for this study using BIGEN, Rudner's (1977c) adaptation of Hamilton and Rovenelli's (1973) procedure. This program uses a Monte Carlo procedure based on Birnbaum's three parameter model to generate hypothetical data with known parameters or characteristics. The three parameters may be manipulated by specifying different mean and standard deviation values for a_g, item discrimination; b_g, item difficulty; and c_g, pseudo-guessing.

In this study only the difficulty parameter was manipulated.

Sixty item hypothetical tests were generated with two levels of mean difficulty, an average of 60% of the items correct and an average of 80% of the items correct. Within each difficulty level three degrees of bias were introduced: 0% of the items biased, 10% of the items biased, and 20% of the items biased. These data sets represent what one might expect to find with standardized, norm-referenced achievement tests and with standardized, criterion-referenced achievement tests. The sixty item test length represents the subscale or subtest lengths of many published tests. The degree of bias covers the range one would expect to find from what is seen in the literature on test bias.

Bias for data generation was defined as a difference of .70 or greater in the area between the ICC's of the two groups. The following formula was used to compute the area between the ICC's (See Rudner 1977c):

$$\phi g = \sum_{-5.000}^{5.000} [\,|\, P(u_g=1|\theta_i) - P'(u_g=1|\theta_i) \,|\,] \,\Delta\theta$$

where $\Delta\theta = .005$

The six methods of analysis listed earlier were employed. Each method was equated by converting all areas or bias estimates to z scores. The value of the area representing an area difference between ICC's of .70 was converted to a z score, too. In this way all values were placed on the same scale of measurement so they could be directly compared.

The item discrimination parameters and guessing parameters for each data set were held constant by assigning the following values:

$a_1 = 1.0$ SD = .05
$a_2 = 1.0$ SD = .05
$c_1 = .05$ SD = .02
$c_2 = .05$ SD = .02

where: a_1 and a_2 are the discrimination parameters for groups 1 and 2 and
c_1 and c_2 are the guessing parameters for groups 1 and 2.

Limits were placed on the item difficulty parameter so that examinees would not exceed 92% correct on an item or fall below 15% correct on an item. Within these constraints the distributions of all item parameters were normal as were the distributions of examine ability.

Methods of Analysis

Six methods were used to assess the generated bias:

1. Three parameter model (PARAM). Data were parameterized using Urry's approach (1970). Areas between ICC's for the two groups were computed using Rudner's (1977c) procedure. The groups were not equated since equating is a linear transformation and would have no effect on the correlation of the area computed between methods. The z scores for the area for each item were also computed.

2. Rasch Model (BICAL). The data was calibrated separately for each group using the BICAL program (Wright, Mead, and Draba, 1978). The area between ICC's was computed using Rudner (1977c) procedure; z scores were computed for each item on the area parameter.

3. Point Biserial (PBIS). Item-total correlations were computed for each group, and the difference between item-total correlations was used as a measure of bias. Z scores for item-total differences were computed.

4. Transformed Item Difficulty (TID). Item difficulty values were computed for each group and transformed to z scores. The distance from a 45° hypothetical regression line through the origin was computed for each item. The distance from the 45° regression line was used as the measure of bias.

5. Chi-Square (CHI). The total score distribution was divided into eight intervals for data sets with average difficulty of 60% of the items correct. For data sets with average difficulty of 80% of the items correct total score distribution was divided into six intervals. The score intervals become column effects and groups were represented as row effects. A standard chi-square for each obtained contingency table was computed. The chi-square was given a sign according to the direction of the difference between the rows. The signed chi-square was then used as a measure of the degree of bias; z scores for these chi-squares were also computed.

6. Factor Analysis (FACTOR). Group membership was included as a dummy variable with the 60 item responses in computing an item intercorrelation matrix. The resulting intercorrelation matrix was reduced with Principal Components Analysis. The factor structure was rotated using varimax rotation. Bias was identified by the factor loadings of the items on the factor which contained the largest factor loading for the group membership variable. Again z scores were computed for the identified bias factor.

Agreement of Methods

The primary measure of the degree of agreement in the identification of biased items with the various methods was the

correlation between the measures of bias for each item. A secondary measure of bias was an examination of the items identified as biased by each of the methods. In an attempt to make these comparisons comparable z score transformations were used to identify the biased items. A biased item was defined as having a z score on the bias measure equal to the z score for an area of .70 in the generated data.

RESULTS

Characteristics of Data Sets

Six data sets were generated, three with an average of 60% of the items correct and three with an average of 80% of the items correct. For each of the two levels of correctness the percentage of biased items was varied in three ways. First, no items were biased; second, 10% of the items were biased, and finally, 20% of the items were biased. This yielded 0 biased items, 6 biased items, and 12 biased items for each level of correctness. Data sets were designated 60-0 for the set with an average difficulty of 60% and no biased items to 80-20 for the set with an average difficulty of 80% and 20% of the items biased; this information is summarized in Table 1.

Table 1

SUMMARY OF DATA SETS GENERATED FOR ANALYSIS

PERCENT OF BIAS	NUMBER OF BIASED ITEMS	NUMBER OF GROUPS	NUMBER OF CASES IN EACH GROUP	AVERAGE PERCENT CORRECT	
				60	80
				DATA SET DESIGNATION	DATA SET DESIGNATION
0	0	2	1000	60- 0	80- 0
10	6	2	1000	60-10	80-10
20	12	2	1000	60-20	80-20

Descriptive statistics for the six sets of computer generated test data are reported in Table 2. Each set closely approximates what would be expected to be found in terms of test characteristics in actual test data. The third and fourth moments computed for each set were similar and indicated that the distributions of total scores were flatter than normal and positively skewed. This was similar to what is observed in actual test data. Cronbach's Alpha, a measure of internal consistency, were similar to those typically obtained with actual data and ranged from .84 to .94.

The average percent correct for total score for the two groups of hypothetical examinees in each data set was kept equal by biasing 50% of the items in favor of one group and 50% in favor of the other group. This resulted in test characteristic curves and total score distributions which were similar. The investigators recognized that in the actual situation the test characteristic curves for the two groups are often different. When this is the case, differences are attributable either to bias or to different abilities in the two groups. In the absence of external criteria, ability and bias are confounded. For this reason data were generated in such a way that test characteristic curves were similar. In this way the intent of the study, to compare methods of detecting test item bias, could be maintained clearly.

Table 2

TEST STATISTICS FOR TWO GROUPS OF 1000 EXAMINEES

DATA SET	GROUP 1			GROUP 2			NUMBER OF BIASED ITEMS
	$\bar{X}$	SD	α	$\bar{X}$	SD	α	
60 - 0	36.99	12.48	.93	36.25	12.87	.94	0
60 - 10	36.08	12.53	.93	35.35	12.84	.94	6
60 - 20	36.01	12.52	.93	36.46	12.40	.93	12
80 - 0	48.15	7.08	.84	48.05	7.61	.86	0
80 - 10	48.52	7.01	.84	47.47	7.64	.86	6
80 - 20	49.17	7.22	.85	47.19	7.57	.85	13

Comparison of Methods

In order to determine how well each of the six methods detected bias, the standard score conversion of the measure of bias for each was correlated with the actual degree of bias generated in the data. The correlations are reported in Table 3.

ICC. Because data were generated by the three parameter ICC model, one would expect substantial correlations between the areas generated with BIGEN and identified by PARAM. Correlations ranged from a low of .75 for the 80-20 condition to a high of .96 for the 60-10 condition. The relatively low correlation obtained for the 80-20 condition might have been due to specific characteristics of the data set; however, this cannot be determined definitely without generating and analyzing other data sets with similar parameters. Correlations between BICAL areas and BIGEN areas ranged from .20 for the 80-20 condition to .96 for the 60-10 condition.

Both PARAM and BICAL correlated well with BIGEN at 60-10, 60-20, and 80-10; at 80-20, however, the correlation dropped markedly. This may be a peculiarity of the specific data set generated for the 80-20 condition, although it should be noted that for the 80-20 set the difference between mean total scores for the two groups was 1.98. None of the other data sets had differences larger than 1.05. Replication by generating other data sets with 80% average difficulty and 20% bias would be required to establish

Table 3

CORRELATIONS BETWEEN THE DEGREE OF BIAS IDENTIFIED BY EACH METHOD AND THE AMOUNT GENERATED WITH BIGEN

	METHOD					
DATA SET	PARAM	BICAL	PBIS	TID	CHI	FACTOR
60 - 10	.96	.96	.07	.96	.86	.10
60 - 20	.94	.98	.36	.97	.91	.95
80 - 10	.87	.95	.54	.95	.75	.92
80 - 20	.75	.20	.61	.96	.75	.95

the reliability of this finding. Varying the difference between group means with additional data sets would indicate whether the magnitude of difference differentially affected the outcome of the analysis.

Point Biserial Correlation. Correlations between PBIS results and BIGEN areas ranged from .07 to .61. These results are consistent with the findings of Rudner (1977) and indicate that, contrasted with ICC methods, PBIS detected biased items poorly.

Transformed Item Difficulties. TID results in terms of distance from the hypothetical regression line and BIGEN areas correlated at or above .95 in all cases. Only the difficulty parameter was varied for generation of these data; thus, the findings were cast into doubt because Hambleton, et al. (1978) and Rudner, Getson, and Knight (1979b) found less favorable results when discrimination and guessing were varied. Another set of data was generated for the 60-10 situation. In this set, variations in discrimination and guessing were introduced; still, the correlation between TID distances and BIGEN areas was .95. Thus, it would appear that TID is not as sensitive to variations in discrimination and guessing as has been reported. Nevertheless, replication of these analyses must be undertaken to provide more definite information on this situation.

Chi-Square. Signed chi-squares correlated with BIGEN areas from 175 to .91. This method, which may be considered a special

instance of ICC, appeared to function well.

Factor Analysis. Correlations between z-scores on the bias factor and BIGEN areas ranged from .10 to .95. Agreement between analysis of the 60-10 data set and the generated data was poor. There seemed to be no logical reason for the low correlation. Factor loadings on the bias factor for the 60-10 set were similar to those on the other sets as were communalities and traces. While the result for the 60-10 set might have been aberrant, additional work must be done to establish the stability of this method.

Accuracy of Identification. Table 4 presents data on the total number of items identified as biased and the number of items correctly identified as biased. Many of the methods consistently over-identify biased items. PARAM, BICAL, TID, and FACTOR all identified more than the number of biased items generated by BIGEN. PBIS over-identified in three instances and underindentified in one instance, while CHI over-identified in two instances and under-identified in two instances.

Each method was differentially successful in identifying biased items correctly. PBIS consistently underidentified the biased items. FACTOR exceeded 80% correct identification in sets 50-20, 80-10, and 80-20; it identified none correctly for set 60-10. TID exceeded 80% correct identification under all four conditions. PARAM and CHI exceeded 80% correct identification under the 60-10 and 60-20 conditions. PARAM exceeded 50% correct

Table 4

COMPARISONS OF NUMBER OF ITEMS IDENTIFIED AS BIASED AND NUMBER OF BIASED ITEMS CORRECTLY IDENTIFIED

	PARAM		BICAL		PBIS		TID		CHI		FACTOR	
	Set 60-10, Number of Biased Items = 6											
	NUMBER	%	NUMBER	%	NUMBER	%	NUMBER	%	NUMBER	%	NUMBER	%
ITEMS IDENTIFIED AS BIASED	9	150	10	167	9	150	9	150	7	117	7	117
IDENTIFIED CORRECTLY	5	83	6	100	2	33	5	83	5	83	0	0
	Set 60-20, Number of Biased Items = 12											
ITEMS IDENTIFIED AS BIASED	16	133	15	125	9	75	18	150	14	117	16	133
IDENTIFIED CORRECTLY	10	83	11	92	5	41	11	92	10	83	11	92
	Set 80-10, Number of Biased Items = 6											
ITEMS IDENTIFIED AS BIASED	9	150	11	183	10	167	10	167	4	67	10	167
IDENTIFIED CORRECTLY	3	50	5	83	4	67	6	100	2	33	5	83
	Set 80-20, Number of Biased Items = 13											
ITEMS IDENTIFIED AS BIASED	16	123	24	184	17	130	15	115	6	46	17	130
IDENTIFIED CORRECTLY	9	69	6	46	8	62	11	85	3	23	12	92

under 80-10 and 80-20, but CHI fell below 50% correct. PBIS did not exceed 70% under any of the conditions. BICAL exceeded 80% for the 60-10, 60-20, and 80-10 conditions but fell below 50% for the 80-20 condition.

CONCLUSIONS AND RECOMMENDATIONS

Summary

This investigation has examined six methods for assessing test item bias. Six data sets simulating varying conditions of difficulty and different degrees of bias were generated with a Monte Carlo technique. Each method was used to analyze each of the data sets. The results were compared, conclusions drawn, and recommendations made.

Conclusions

Overall, the method which had consistently the highest correlations with BIGEN and identified correctly the largest percentage of biased items was TID. It was followed by PARAM, then, by BICAL, next by FACTOR, and last by CHI. PBIS functioned erratically.

A comparison of the six methods is shown in Table 5. The only method functioning poorly is Point Biserial. The basis for this judgment is the erratic performance of the method and the arbitrariness of the bias criterion. The method which appeared to function best, which worked for situations with smaller numbers of examinees in each group, and which was simple to compute and interpret is Transformed Item Difficulty. Item Characteristic Curve approaches fared less well in this study than they have in others; however, they can provide useful information and reasonable accuracy. ICC's seem best used when large numbers of examinees are involved, when computational resources and statistical experts are available, and when additional analyses are being attempted. Chi-square appears to be appropriate when relatively large numbers of examinees (100 per group) are available, when a direct test of the significance of difference is desirable, and when a straightforward definition and explanation is desirable. Factor Analysis fared better in the present study than it has in others. Still, it is probably most appropriate as a research tool for investigators who have computational and statistical resources beyond those usually available to the practitioner.

Table 5

COMPARISON OF METHODS

METHOD	DEFINITION OF BIAS	PARAMETERS EXAMINED	COMPUTATIONAL EASE	EASE OF INTERPRETATION
PARAM	Different Probability of Correct Response Given Equal Ability	Item Discrimination Item Difficulty Guessing	Computer Necessary	Simple
BICAL	Different Probability of Correct Response Given Equal Ability	Item Difficulty	Computer Necessary	Simple
PBIS	Difference Between Magnitudes of Correlations Exceed an Arbitrary Value	Item-Total Correlation	Hand Calculation Possible Computer Desirable	Simple
TID	Item by Group Interaction	Item Difficulty	Hand Calculation Possible Computer Desirable	Simple
CHI	Different Proportions of Examinees Attaining an Item Correct Within an Ability Level	Item Difficulty	Hand Calculation Possible Computer Desirable	Simple
FACTOR	Major Factor Loadings Bias Factor	Item Intercorrelation	Computer Only	Moderately Difficult

Table 5 (Continued)

COMPARISON OF METHODS

METHOD	MEASURE OF BIAS	AVAILABILITY OF BIAS CRITERION	NUMBER OF EXAMINEES NEEDED PER GROUP	ESTIMATE OF FUNCTION
PARAM	Area Between ICC Curves	Significance Test of Magnitude of Areas	At Least 10 Times the Number of Items; 1000 per Group Most Desirable	Good
BICAL	Area Between ICC Curves	Significance Test of Magnitude of Areas	At Least 10 Times the Number of Items; 1000 per Group Most Desirable	Good
PBIS	Magnitude of Differences Exceed a Specified Amount	Arbitrary Designation	50 to 100 per Group	Poor
TID	Distance from Regression Line	Arbitrary Designation	30 to 50 per Group	Excellent
CHI	Magnitude of Chi-Square Value	Significance of Chi-Square	100 or more per Group	Moderate
FACTOR	Magnitude of Factor Loading	None	At Least 10 times the Number of Items; 1000 per Group Most Desirable	Moderate

REFERENCES

Angoff, W. H. A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, May 1972.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-105.

Angoff, W. H., & Modu, C. C. Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test. New York: College Entrance Examination Board, 1973.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Ma: Addison-Wesley, 1968, Chapters 17-20.

Bourne, C. A psychometric investigation of test bias. Unpublished masters thesis. Sacramento: California State University, Sacramento, 1977.

Durovic, J. Definitions of test bias: a taxonomy and an illustration of an alternative model. Unpublished doctoral dissertation. Albany: State University of New York, 1975

Echternacht, B. A quick method of determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Green, D. R. Racial and ethnic bias in test construction. USOE Contract No. OEC-9-70-58 (0057), CTB/McGraw-Hill, Monterey, 1971.

Green, D. R. Reducing bias in achievement tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Green, D. R., & Draper, J. F. Exploratory studies of bias in achievement tests. Monterey: CTB/McGraw-Hill, 1972.

Groome, M. L. A comparison of two approaches for item bias identification. Unpublished masters thesis. Sacramento: California State University, Sacramento, 1977.

Groome, W. R. Item bias: a comparison of two approaches. Unpublished masters thesis. Sacramento: California State University, Sacramento, 1977.

Hambleton, R. K., & Rovanelli, R. A. Fortran IV program for generating item response data from logistic test models. Behavioral Science, 1973, 18, 74.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hicks, M. M., Donlon, T. F., & Wallmark, M. M. Sex differences in item responses on the Graduate Record Examination. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, April 1976.

Ironson, G. A comparative analysis of several methods of assessing item bias. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Jensen, A. P. An examination of culture bias in the Wonderlic Personnel Test. Arlington, VA: Eric Clearinghouse, 1973, (ERIC Document Reproduction Service EO, 86 726).

Lord, F. M. A theory of test scores: Psychometric monograph No. 7. Princeton: Educational Testing Service, 1952.

Lord, F. M. A study of item bias using item characteristic curve theory. Proceedings of the Third Congress of Cross-Cultural Psychology, Tilburg, Holland, 1977.

Merz, W. R. Factor analysis as a technique in analyzing test item bias. Paper presented at the annual meeting of the California Educational Research Association, Los Angeles, 1973.

Merz, W. R. Test fairness and test bias: a review of procedures. In M. J. Wargo & D. R. Green, Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey: CTB/McGraw-Hill, 1978.

Merz, W. R., & Rudner, L. M. Bias in testing: a presentation of selected methods. A paper as part of a co-sponsored symposium, Empirical Evaluations of Various Approaches for Identifying Biased Test Item(s), at the annual meeting of the American Educational Research Association and the National Council for Measurement in Education, Toronto, March 1978.

Ozenne, D. G., Van Gelder, N. C. & Cohen, A. J. Emergency school aid act (ESAA) national evaluation, achievement test restandardization. Santa Monica: Systems Development Corporation, 1974.

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

Pine, S. M. Applications of item characteristic curve theory to the problems of test bias. In D. J. Weiss (ed.). Applications of computerized adaptive testing, (RR 77-1). Minneapolis: University of Minnesota Psychometric Methods Program, March 1977.

Potthoff, R. F. Statistical aspects of the problems of biases in psychological test. University of North Carolina Institute of Statistics Mimeo Services, No. 479, 1966.

Rasch, G. Probablistic models for intelligence and attainment tests. Copenhagen: Nilsen and Lydiche (for Denmark's Paedagogich Institute), 1960.

Rudner, L. M. An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977a.

Rudner, L. M. A closer look at latent trait parameter invariance. Paper presented at the annual meeting of the New England Educational Research Organization, Manchester, NH, May 1977b.

Rudner, L. M. An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation, The Catholic University of America, 1977c.

Rudner, L. M. Efforts toward the development of unbiased selection and assessment instruments. Paper presented at the Third Annual Symposium on Educational Testing, University of Leyden, the Netherlands, 1977d.

Rudner, L. M. Using standard tests with the hearing impaired: the problems of item bias. Volta Review, 1978, 80, 31-40.

Rudner, L. M., Getson, P. R., & Knight, D. L. Biased item detection techniques. Gallaudet College, Washington, D.C., 1979a.

Rudner, L. M., Getson, P. R., & Knight, D. L. The effect of various test and item properties on five approaches to biased item detection. Paper presented at the annual meeting of the National Council on Educational Measurement, San Francisco, April 1979b.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C., April 1975.

Scheuneman, J. A procedure for evaluating item bias in the absence of an outside criterion. Paper presented at the annual meeting of the American Educational Research Association, April 1976.

Strassberg-Rosenberg, B., & Donlon, R. F. Context influences on sex differences in performance and aptitude tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D. C., 1975.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1971.

Thorndike, R. L. Reliability. In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951.

Tucker, L. R. An interbattery method of factor analysis. Psychometrika, 1958, 23, 111-136.

Urry, V. W. A Monte Carlo investigation of logistic mental test models. Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V. W. Ancillary estimators for the parameters of mental test models. Paper presented at the American Psychological Association Convention, Chicago, August 1975.

Weiss, D. J. (Ed.). Applications of computerized adaptive testing, (RR 77-1). Minneapolis: University of Minnesota Psychometric Methods Program, March 1977.

Wingersky, M. S., & Lork, F. M. A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses, (RM 73-2). Princeton: Educational Testing Service, 1973.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., Mead, R. J., & Draba, R. Detecting and correcting test item bias with a logistic response model. Research Memorandum No. 22, Statistical Laboratory, Department of Education, University of Chicago, 1976.